Submitted to the Annals of Applied Statistics

DEEP LEARNING OF SEMI-COMPETING RISK DATA VIA A NEW NEURAL EXPECTATION-MAXIMIZATION ALGORITHM

BY STEPHEN SALERNO [1,a] AND ZHILIN ZHANG [2,b] AND YI LI [2,c]

[1] *Public Health Sciences Division, Biostatistics, Fred Hutchinson Cancer Center, Seattle,* [a] *ssalerno@fredhutch.org*

[2] *Department of Biostatistics, University of Michigan, Ann Arbor,* [b] *zzhilin@umich.edu;* [c] *yili@umich.edu*

Prognostication for lung cancer, a leading cause of mortality, remains a complex task, as it needs to quantify the associations of risk factors and health events spanning a patient's entire life. One challenge is that an individual's disease course involves non-terminal (e.g., disease progression) and terminal (e.g., death) events, which form *semi-competing* relationships. Our motivation comes from the Boston Lung Cancer Study, a large lung cancer survival cohort, which investigates how risk factors influence a patient's disease trajectory. Following developments in the prediction of time-to-event outcomes with neural networks, deep learning has become a focal area for the development of risk prediction methods in survival analysis. However, limited work has been done to predict multi-state or semi-competing risk outcomes, where a patient may experience adverse events such as disease progression prior to death. We propose a neural expectation-maximization algorithm for semi-competing risks to bridge the gap between classical semi-competing survival models and deep learning. Our algorithm enables estimation of the nonparametric baseline hazards of each state transition, risk functions of predictors, and the degree of dependence among different transitions, via a multitask deep neural network with transition-specific sub-architectures. We apply our method to the Boston Lung Cancer Study and investigate the impact of clinical and genetic predictors on disease progression and mortality.

1. Background. Lung cancer remains a leading cause of cancer mortality, with a 5-year survival rate below 20% worldwide (Bade and Cruz, 2020). Prognosis is difficult to predict, as it depends on a lifetime of risk factors and health events (Goel et al., 2021). Moreover, key 'non-terminal' events such as disease recurrence and progression are often used to guide care, including the availability of second-line treatment options, but form *semi-competing* relationships with mortality, meaning that death can censor these events but not vice versa (Fine, Jiang and Chappell, 2001). Similar semi-competing dynamics occur with death in other chronic diseases, including treatment-limiting events such as radiation toxicity in patients with HPV-related cancers and volume overload, vascular access failure, or graft failure (post-transplantation) in patients with end-stage renal disease (Shu et al., 2018; Haddad et al., 2008). Ignoring this dependence is known to bias inference and prediction (Jazić et al., 2016). Prognostic heterogeneity also arises from complex and interacting factors such as smoking status, genetic mutations, demographics, and comorbid conditions, making individualized risk prediction challenging (Ashworth et al., 2014; Gaspar et al., 2012).

Deep learning has sparked interest in the advancement of risk prediction methods within the field of survival analysis (Faraggi and Simon, 1995; Katzman et al., 2016; Ranganath et al., 2016; Jing and Smola, 2017; Kvamme, Borgan and Scheel, 2019; Hao et al., 2021). Many of these approaches extend the Cox proportional hazards model (Cox, 1972) to nonlinear predictions or use a patient's survival status directly as a binary training label, predicting a patient's survival probability. More recently, competing risk and multi-state models extend

Keywords and phrases: Deep Learning, Semi-Competing Risks, EM Algorithm, Brier Score, Lung Cancer.

these methods to settings where multiple event types mutually censor one another (Lee et al., 2018; Lee, Yoon and Van Der Schaar, 2019; Aastha and Liu, 2020; Tjandra, He and Wiens, 2021). Such methods characterize the risk of one or more competing events by estimating either the cause-specific or subdistribution hazards of each event type. However, these do not accommodate the joint prediction of correlated events or the study of outcome trajectories between events, important considerations for semi-competing risks.

Several recent works have developed methods for risk prediction with semi-competing outcomes. Xu, Kalbfleisch and Tai (2010) proposed an approach based on the illness-death model, which was defined by the hazards of transitioning between disease states. The authors used a shared gamma frailty conditional Markov model, parameterized by three Cox-based hazard functions, and a semiparametric maximum likelihood estimation (MLE) approach. The gamma frailty, a type of subject-specific random effect, captures the strength of unobserved, individual-level heterogeneity that drives dependence between the illness and death transitions (Li et al., 2020). Estimating its variance allows us to quantify latent risk subgroups, assess how much of the observed transition dynamics are driven by shared frailties versus measured covariates, and thus directly informs prognostic stratification and personalized treatment planning. Lee et al. (2015) also adopted the gamma frailty formulation of the illness-death model with Cox-type hazards, but instead proposed a semiparametric Bayesian approach for estimation. Lee, Rondeau and Haneuse (2017) formulated Bayesian (semi)parametric approaches with an illness-death accelerated failure time (AFT) model, adopting an additive normal frailty, rather than a multiplicative gamma frailty. Lee, Gilsanz and Haneuse (2021) further proposed a spline-based approach, for additional flexibility.

More recently, Gorfine et al. (2021) developed a Cox-based marginalized gamma frailty illness-death model and estimated it using a semiparametric pseudo-likelihood, and Kats and Gorfine (2022) proposed an AFT-based gamma frailty model via semiparametric MLE. In addition, approaches such as the one proposed by Jiang and Haneuse (2017) consider transformation illness-death models with parametric error distributions, but nonparametric frailty distributions. In high dimensions, Reeder, Lu and Haneuse (2022) proposed a regularized estimation approach which combines a non-convex and structured fusion penalization. Salerno and Li (2022) developed a deep learning framework for predicting semi-competing risk outcomes based on the model of Xu, Kalbfleisch and Tai (2010), with parametric baseline hazards estimated via gradient methods. However, it remains challenging to estimate nonparametric baseline hazards, which confer greater robustness, within a deep learning framework.

To address this, we propose a neural expectation-maximization algorithm for semi-competing risks (NEM-SCR), which bridges the gap between classical semi-competing survival models and deep learning. Our proposal is to replace the traditional parametric M-step for frailty-based illness-death models with two non-parametric updates. In our M-step, we exploit closed-form non-parametric maximum likelihood estimates to recover the baseline hazards of transitioning between model states. In our N-step, we fit a multi-head neural network via stochastic gradient descent to learn flexible covariate-dependent risk functions. Because the semi-competing illness-death model's likelihood decomposes additively across state transitions, each network head optimizes a transition-specific loss function. The NEM-SCR algorithm enables estimation of nonparametric baseline hazards, nonparametric risk functions of our predictors, and the degree of dependence between events. Supplemental Table F.1 provides a comparison of key modeling assumptions and methodological characteristics, including model structure, flexibility, computational features, and data requirements, between our model and classical semi-competing models. Section 2 motivates this work, Section 3 reviews the illness-death model, and Section 4 details our proposed algorithm. Section 5 introduces new frameworks for evaluating predictive performance in this setting by extending the widely-used Brier score and concordance index to the bivariate survival function. In Section 6, we assess the performance of our method in simulation before applying it to the Boston Lung Cancer Study in Section 7. We conclude with a discussion and future directions.

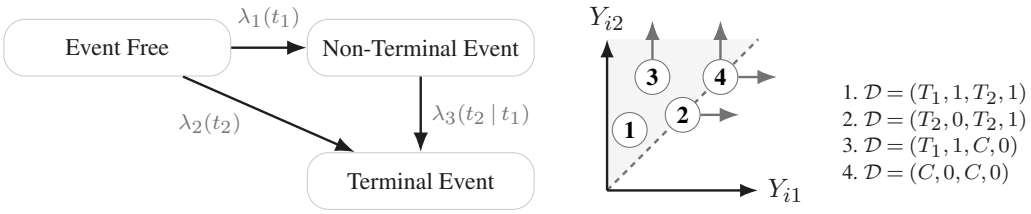

Fig 1: Graphical representations of semi-competing risks data. Left: Illness-death model with three state transition rates characterized by $\lambda_1(t_1)$, $\lambda_2(t_2)$, and $\lambda_3(t_2 \mid t_1)$. Right: Observable space for semi-competing data with example observations: (1) both events observed, (2) terminal event observed, (3) non-terminal event observed, and (4) neither event observed. Arrows represent the direction of censoring, and $\mathcal{D} = (Y_1, \delta_1, Y_2, \delta_2)$ represents the data under each example observation.

2. The Boston Lung Cancer Study. We are motivated by the Boston Lung Cancer Study (BLCS), one of the largest global lung cancer survival cohorts (Christiani, 2017). A key objective of the BLCS is to understand how clinical and lifestyle risk factors, as well as adverse events such as disease progression, affect survival. Accurate prediction of both non-terminal (progression) and terminal (death) events enables dynamic clinical decisions regarding surveillance, restaging scans, and adjuvant or salvage therapy (Orstad et al., 2023). Since disease progression often precedes death (Inamura and Ishikawa, 2010), quantifying their dependence is crucial for personalized risk stratification and patient counseling.

The BLCS cohort comprises patients from Massachusetts General Hospital and Dana-Farber Cancer Institute, with detailed data on demographics, pathology, treatments, and oncogenic mutations (Lynch et al., 2004; Paez et al., 2004). Patients are enrolled at diagnosis and followed until death, with disease recurrence or progression recorded as key non-terminal events. For early-stage (I-IIIa) patients, the non-terminal event is time to recurrence; for late-stage patients, it is time to progression. These are determined radiographically using RECIST guidelines (Eisenhauer et al., 2009). Mortality data are obtained from the BLCS registry and the National Death Index. In our analysis, the semi-competing events of interest are cancer progression and death, subject to censoring at the end of follow-up.

3. Notation. Let T_1 and T_2 denote the times from an initial event-free state (e.g., diagnosis) to a non-terminal event (recurrence or progression) and terminal event (death), respectively, where the non-terminal event is subject to censoring by death. Under the illness-death framework (Andersen et al., 2012), we model transitions among three states: event-free, non-terminal, and terminal. As shown in the left panel of Figure 1, $\lambda_1(t_1)$ denotes the hazard from event-free to non-terminal at $t_1 > 0$, $\lambda_2(t_2)$ the hazard from event-free to terminal at $t_2 > 0$ (without prior non-terminal event), and $\lambda_3(t_2 \mid t_1)$ the hazard from non-terminal to terminal at $t_2 > t_1$. These transition hazards are specified as

$$\lambda_1(t_1) = \lim_{\Delta \to 0} \Pr[T_1 \in [t_1, t_1 + \Delta) \mid T_1 \geq t_1, T_2 \geq t_1]/\Delta, \tag{3.1}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} \Pr[T_2 \in [t_2, t_2 + \Delta) \mid T_1 \geq t_2, T_2 \geq t_2]/\Delta, \tag{3.2}$$

$$\lambda_3(t_2 \mid t_1) = \begin{cases} \lim_{\Delta \to 0} \Pr[T_2 \in [t_2, t_2 + \Delta) \mid T_1 = t_1, T_2 \geq t_2]/\Delta, & t_2 > t_1 > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

Similar formulations have been proposed by Xu, Kalbfleisch and Tai (2010); Gorfine et al. (2021); Kats and Gorfine (2022), and others. Both the non-terminal and terminal events can be subject to independent censoring; we focus only on the case of right censoring, whereby a subject may be lost to follow-up or the study ends before the event has occurred. For the ith

individual in a sample of n subjects, we denote the censoring time by C_i and add a subscript i to T_1, T_2 for this individual. The observed data are denoted as

$$\mathcal{D} = \{(Y_{i1},\ \delta_{i1},\ Y_{i2},\ \delta_{i2});\ i = 1, \ldots, n\},$$

where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \leq Y_{i2})$, and $I(\cdot)$ denotes the indicator function. Note that our observable data take on probability mass only in the *upper wedge* on which $Y_{i1} \leq Y_{i2}$ and arise from four potential cases: (1) a subject experiences both event types, (2) a subject experiences only the terminal event, (3) a subject experiences only the non-terminal event, or (4) a subject experiences neither event prior to the end of follow up (Figure 1, Right Panel). Following the original work of Xu, Kalbfleisch and Tai (2010), we model (3.1) - (3.3) by extending a Cox-type hazard function for each state transition to include a baseline hazard, a frailty term, and a patient's covariates as

$$\lambda_1(t_1 \mid \gamma_i, \boldsymbol{x}_i) = \gamma_i \lambda_{01}(t_1) \exp\{h_1(\boldsymbol{x}_i)\}, \tag{3.4}$$

$$\lambda_2(t_2 \mid \gamma_i, \boldsymbol{x}_i) = \gamma_i \lambda_{02}(t_2) \exp\{h_2(\boldsymbol{x}_i)\}, \tag{3.5}$$

$$\lambda_3(t_2 \mid t_1, \gamma_i, \boldsymbol{x}_i) = \begin{cases} \gamma_i \lambda_{03}(t_2) \exp\{h_3(\boldsymbol{x}_i)\}, & t_2 > t_1 > 0; \\ 0, & \text{otherwise}, \end{cases} \tag{3.6}$$

where γ_i is a subject-specific random effect, termed *frailty*, that induces dependence among the three transition processes, $\lambda_{01}(t_1)$, $\lambda_{02}(t_2)$, and $\lambda_{03}(t_2)$ are the baseline hazard functions for the three state transitions, respectively, $\boldsymbol{x}_i$ is a p-vector of clinically relevant, time-independent predictors such as patient socio-demographic status, medical history data, and comorbid conditions collected at baseline, and $h_g(\boldsymbol{x}_i), g \in \{1, 2, 3\}$, are log-risk functions which relate a patient's covariates to the hazard rates for each potential transition. Unlike existing approaches, we do not parameterize $h_g(\boldsymbol{x}_i)$, but estimate these functions nonparametrically through neural network architectures. For identifiability, we fix output-layer biases to zero and assume individual frailties $\gamma_i \sim \text{Gamma}(1/\theta,\ 1/\theta)$ with $\mathbb{E}(\gamma_i) = 1$ and $\text{Var}(\gamma_i) = \theta$. The Gamma frailty is widely used (Xu, Kalbfleisch and Tai, 2010; Haneuse and Lee, 2016; Kats and Gorfine, 2022), though unverifiable within the bivariate distribution's lower wedge (Figure 1), a general limitation of semi-competing risk models. Alternative specifications such as finite mixtures have been proposed (Gasperoni et al., 2020; Chee et al., 2021). We adopt the frailty formulation given its clinical relevance to the BLCS study, however if the dependence parameter, θ, were considered a nuisance, the marginalized model of Gorfine et al. (2021) could be used instead. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\Lambda_{0g}(t) = \int_0^t \lambda_{0g}(u)\, du$ for $g = 1, 2, 3$, and $\boldsymbol{\psi} = \{\Lambda_{01}, \Lambda_{02}, \Lambda_{03}, h_1, h_2, h_3, \theta\}$. The complete-data likelihood is

$$\begin{aligned} L(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} & \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\lambda_{01}(Y_{i1})\, e^{h_1(\boldsymbol{x}_i)}\right]^{\delta_{i1}} \\ & \times \left[\lambda_{02}(Y_{i2})\, e^{h_2(\boldsymbol{x}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\lambda_{03}(Y_{i2})\, e^{h_3(\boldsymbol{x}_i)}\right]^{\delta_{i1}\delta_{i2}} \\ & \times \exp\Big\{-\gamma_i \Big[\Lambda_{01}(Y_{i1})\, e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}(Y_{i1})\, e^{h_2(\boldsymbol{x}_i)} \\ & \quad + \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] e^{h_3(\boldsymbol{x}_i)}\Big]\Big\}. \end{aligned} \tag{3.7}$$

See Supplement A for the derivation of (3.7). One could integrate out γ_i in (3.7) and maximize the resulting likelihood, but two challenges arise: (1) the integral lacks a closed form,

and (2) the likelihood has no maximizer over the space of absolutely continuous baseline hazards (Johansen, 1983). To address these, we develop an EM-type algorithm treating γ_i as latent variables and iteratively updating parameters. We further restrict the cumulative baseline hazards Λ_{01}, Λ_{02}, and Λ_{03} to piecewise constant, right-continuous (CADLAG) functions with jumps at observed event times. The resulting maximizers are the nonparametric maximum likelihood estimates (NPMLEs), which remain agnostic to the true baseline hazard form. We then modify (3.7) by replacing $\lambda_{0g}(t)$ with the jump size $\Delta\Lambda_{0g}(t) = \Lambda_{0g}(t) - \Lambda_{0g}(t-)$ (Li and Lin, 2000; Kim et al., 2012), yielding

$$
\begin{aligned}
\tilde{L}\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right) = \prod_{i=1}^{n} & \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\Delta\Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)}\right]^{\delta_{i1}} \\
& \times \left[\Delta\Lambda_{02}\left(Y_{i2}\right) e^{h_2(\boldsymbol{x}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\Delta\Lambda_{03}\left(Y_{i2}\right) e^{h_3(\boldsymbol{x}_i)}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} \right.\right. \\
& \left.\left. + \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] e^{h_3(\boldsymbol{x}_i)}\right]\right\}.
\end{aligned} \tag{3.8}
$$

This 'complete' likelihood (3.8), defined on a step-wise constant space for Λ_{0g}, will serve as the basis of our proposed algorithm, as detailed below.

4. A Neural Expectation-Maximization Algorithm for Semi-Competing Risks. One challenge stands out when we design our algorithm, as the risk functions, h_g, are completely nonparametric. Therefore, we propose to incorporate a deep learning step into the M-step, which maximizes the expected log conditional likelihood given the data and the current parameter estimates. We term this version of a "neural expectation-maximization algorithm" as a neural EM algorithm for semi-competing risks (NEM-SCR). Specifically, by viewing the frailty term as a missing variable, the algorithm iterates between an expectation (E), a maximization (M) and a neural (N; i.e., deep learning) step. In the E-step, we compute the conditional expectation of the log-likelihood (3.8) given the observed data, $\mathcal{D}$, and the current estimates of $\boldsymbol{\psi}$, denoted by $\boldsymbol{\psi}_c$, wherein the conditional expectation is with respect to the distribution of $\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c$. In the M-step, we calculate the NPMLEs of Λ_{0g}, while assuming known values of h_g. Subsequently, we substitute these estimates into the conditional expectation of the log-likelihood (3.8). This process yields the conditional expectation of the log 'profile' likelihood, serving as the objective function for utilizing deep neural networks to derive estimates for the log risk functions and frailty variance. By 'profile likelihood,' we are referring to a technique of maximum likelihood estimation in the presence of nuisance parameters. That is, for a joint likelihood $L(\theta, \eta)$, the 'profile likelihood' of θ is defined as $L_P(\theta) = \max_\eta L(\theta, \eta)$, where taking the maximum over η is termed 'profiling out' η. We use a profile likelihood because the deep neural network loss functions in our N-step depend on the nonparametric maximum likelihood estimates of the baseline hazards from the M-step.

4.1. *Conditional Frailty Distribution.* It follows that the conditional distribution of γ_i, given $\mathcal{D}$ and $\boldsymbol{\psi}$, is Gamma$(\tilde{a}, \tilde{b})$, where

$$
\begin{aligned}
\tilde{a} &= \frac{1}{\theta} + \delta_{i1} + \delta_{i2}, \\
\tilde{b} &= \frac{1}{\theta} + \Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)} + \delta_{i1}\left\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\right\} e^{h_3(\boldsymbol{x}_i)}.
\end{aligned}
$$

Then, $\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}] = \tilde{a}/\tilde{b}$, and $\mathbb{E}[\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}] = \digamma(\tilde{a}) - \log(\tilde{b})$, where $\digamma(\tilde{a}) = \partial \log[\Gamma(\tilde{a})]/\partial\tilde{a}$ and $\Gamma(\cdot)$ is the gamma function. Both quantities, with $\boldsymbol{\psi}$ replaced by $\boldsymbol{\psi}_c$, are needed for the E-Step. See Supplement B for the full derivation of the conditional frailty distribution.

4.2. *E-Step.* The E-step calculates the expected log-conditional likelihood of the 'complete' data given the observed data and the current estimate of parameters, i.e., $\mathcal{D}, \boldsymbol{\psi}_c$:

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}_c) = \mathbb{E}\left[\log \tilde{L}(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] = Q_1 + Q_2 + Q_3 + Q_4, \tag{4.1}$$

where we recall $\boldsymbol{\psi} = (\Lambda_{0g}, h_g, \theta); g = 1,2,3$ represents the unknown parameters to be estimated, or updated, and Q_1, Q_2, Q_3, and Q_4 are the additive pieces of the 'Q' function, each involving non-overlapping unknown parameters:

$$\begin{aligned}
Q_1 &= \sum_{i=1}^{n} \delta_{i1}\mathbb{E}[\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}_c] + \delta_{i1}\left\{\log\left[\Delta\Lambda_{01}(Y_{i1})\right] + h_1(\boldsymbol{x}_i)\right\} \\
&\quad - \mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]\Lambda_{01}(Y_{i1})\, e^{h_1(\boldsymbol{x}_i)}, \\
Q_2 &= \sum_{i=1}^{n} \delta_{i2}\mathbb{E}[\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}_c] + (1-\delta_{i1})\,\delta_{i2}\left\{\log\left[\Delta\Lambda_{02}(Y_{i2})\right] + h_2(\boldsymbol{x}_i)\right\} \\
&\quad - \mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]\Lambda_{02}(Y_{i1})\, e^{h_2(\boldsymbol{x}_i)}, \\
Q_3 &= \sum_{i=1}^{n} \delta_{i1}\delta_{i2}\left\{\log\left[\Delta\Lambda_{03}(Y_{i2})\right] + h_3(\boldsymbol{x}_i)\right\} \\
&\quad - \mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]\delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}I(Y_{i1})]e^{h_3(\boldsymbol{x}_i)}, \\
Q_4 &= \sum_{i=1}^{n} -\frac{1}{\theta}\log(\theta) + \left(\frac{1}{\theta} - 1\right)\mathbb{E}[\log(\gamma_i)|\mathcal{D},\boldsymbol{\psi}_c] - \frac{1}{\theta}\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c] - \log\Gamma\left(\frac{1}{\theta}\right).
\end{aligned}$$

4.3. *M-Step.* In the M-step, we maximize (4.1) with respect to $\boldsymbol{\psi}$ to obtain its updated values. The separability of $Q_1, \ldots, Q_4$ allows us to estimate Λ_{0g}, h_g by maximizing the $Q_g; g = 1,2,3$, respectively, and estimate θ by maximizing Q_4. As Λ_{0g}, h_g are nonparametric, we adopt a profiling approach to facilitate maximization. For each $g = 1,2,3$, we maximize Q_g with respect to the jump sizes of Λ_{0g}, fixing h_g. This yields Breslow-type estimates:

$$\begin{aligned}
\widehat{\Delta\Lambda_{01}}(t) &= \frac{\sum_{i=1}^{n}\delta_{i1}I\left[Y_{i1}=t\right]}{\sum_{i=1}^{n}\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]I\left[Y_{i1}\geq t\right]\exp\left\{h_1(\boldsymbol{x}_i)\right\}}, \\
\widehat{\Delta\Lambda_{02}}(t) &= \frac{\sum_{i=1}^{n}(1-\delta_{i1})\,\delta_{i2}I\left[Y_{i2}=t\right]}{\sum_{i=1}^{n}\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]I[Y_{i1}\geq t]\exp\left\{h_2(\boldsymbol{x}_i)\right\}}, \\
\widehat{\Delta\Lambda_{03}}(t) &= \frac{\sum_{i=1}^{n}\delta_{i1}\delta_{i2}I\left[Y_{i2}=t\right]}{\sum_{i=1}^{n}\mathbb{E}[\gamma_i|\mathcal{D},\boldsymbol{\psi}_c]\delta_{i1}\left[I(Y_{i2}\geq t) - I(Y_{i1}\geq t)\right]\exp\left\{h_3(\boldsymbol{x}_i)\right\}}.
\end{aligned}$$

See the detailed derivation in Appendix B.

4.4. *N-Step.* From the M-step, we have estimates $\widehat{\Delta\Lambda_{0g}}(t)$ and $\widehat{\Lambda_{0g}}(t) = \sum_{s\leq t}\widehat{\Delta\Lambda_{0g}}(s)$ for $g = 1,2,3$. Plugging these estimates into Q_1, Q_2, and Q_3 yields the expected log-*profile* likelihood for h_1, h_2, h_3, respectively (up to additive constants; see Appendix B). That is, with an added subscript P (for profile), we have that

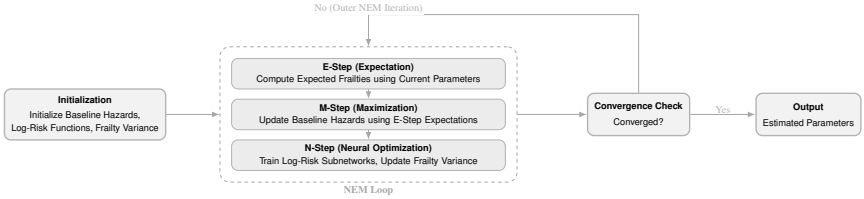

Fig 2: Overview of the neural expectation-maximization algorithm for semi-competing risks.

$$Q_{1,P} = \sum_{i=1}^{n} \delta_{i1} \left\{ h_1(\boldsymbol{x_i}) - \log \left[\sum_{j=1}^{n} \mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\left(Y_{j1} \geq Y_{i1}\right) e^{h_1(\boldsymbol{x}_j)} \right] \right\},$$

$$Q_{2,P} = \sum_{i=1}^{n} (1-\delta_{i1})\delta_{i2} \left\{ h_2(\boldsymbol{x_i}) - \log \left[\sum_{j=1}^{n} \mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\left(Y_{j2} \geq Y_{i2}\right) e^{h_2(\boldsymbol{x}_j)} \right] \right\},$$

$$Q_{3,P} = \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \left\{ h_3(\boldsymbol{x_i}) - \log \left[\sum_{j=1}^{n} \mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right] \delta_{j1} I\left[Y_{j2} \geq \max(Y_{i2}, Y_{j1})\right] e^{h_3(\boldsymbol{x}_j)} \right] \right\}.$$

The functions above resemble the Cox partial likelihood. In the N-step, we treat each profile objective $Q_{g,P}$ $(g = 1, 2, 3)$ as a distinct 'head' in a multitask neural network. Each subject's covariates, $\boldsymbol{x}_i$, are passed through three subnetworks that output log-risk estimates, $\hat{h}_g(\boldsymbol{x}_i)$. Each subnetwork is a fully connected feedforward network with L layers and k_l neurons per layer, applying nonlinear activations in hidden layers (e.g., ReLU) and a linear activation in the output layer (Yegnanarayana, 2009). For identifiability, we impose $h_g(\mathbf{0}) = 0$ by setting output layer biases to zero. Each subnetwork maximizes its corresponding profile likelihood using mini-batch stochastic gradient descent or Adam (Kingma and Ba, 2014). At each iteration, subjects are randomly permuted and divided into batches of size B. A forward pass computes $\hat{h}_g$, and automatic differentiation (Paszke et al., 2017) yields updates:

$$\boldsymbol{W}_l \leftarrow \boldsymbol{W}_l - \eta \nabla_{\boldsymbol{W}_l}, \quad b_l \leftarrow b_l - \eta \nabla_{b_l},$$

where η is the learning rate. Dropout and ℓ_2 regularization are applied to mitigate overfitting. Hyperparameters (e.g., the number of layers, nodes per layer, dropout rate, regularization, and learning rate) are tuned by grid search. Training continues until the change in expected log-profile likelihood is below a pre-specified tolerance.

The frailty variance, θ, enters only through Q_4, which we optimize separately via an inner Adam loop with early stopping. This decoupling simplifies computation and stabilizes estimation. After convergence, $\hat{h}_g$ and $\hat{\theta}$ are updated, and together with the estimated baseline hazards from the M-step, are fed back into the E-step. The NEM-SCR procedure iterates until overall convergence. Initialization uses Nelson-Aalen estimates for Λ_{0g} and Weibull-Cox estimates (via `SemiCompRisks`; see Alvares et al., 2019) for h_g and θ. Figure 2 above and Algorithm 1 in Supplement B summarize the full NEM-SCR algorithm. The implementation, developed in PyTorch (Stevens, Antiga and Viehmann, 2020), is available at https://github.com/salernos/SemiCompDNN.

4.5. *Relation to Prior Neural EM Frameworks.* There is a growing body of literature that proposes hybrid EM algorithms with deep architectures for latent variable modeling, including variational deep embedding (Jiang et al., 2016), unsupervised deep embedding

(Xie, Girshick and Farhadi, 2016), and the neural expectation maximization algorithm of Greff, Van Steenkiste and Schmidhuber (2017). Subsequent extensions have explored relational reasoning (Van Steenkiste et al., 2018), noisy multi-label text classification (Chen et al., 2022), and bridging inference gaps in neural processes (Wang, Federici and van Hoof, 2025). Earlier iterative methods such as those proposed by Carreira-Perpinan and Hinton (2005) and Ba and Caruana (2014) similarly adopt EM-style parameter refinement within deep architectures. These types of models typically introduce latent variables that represent low-dimensional feature embeddings or mixtures assignments, and they are optimized with respect to a variational or amortized evidence lower bound (ELBO). In contrast, we note that our latent variable is a subject-specific frailty that induces dependence across the semi-competing event transitions. The proposed algorithm retains an explicit E-step and a closed-form M-step update for the nonparametric baseline hazards, but replaces the M-step update to the risk functions with deep architectures. This preserves the statistical interpretability of the semi-competing survival model, while incorporating the flexibility of deep learning.

5. Measures of Predictive Performance.

5.1. *Bivariate Brier Score.* To assess predictive performance in semi-competing risk settings, we propose a bivariate extension to the inverse probability of censoring weighting (IPCW)-approximated Brier Score (Brier et al., 1950). Let $S_i(t) = \Pr(T_{i1} > t,\ T_{i2} > t)$ denote the disease-free survival function for individual i at a fixed time point, t. Further, denote an estimate of $S_i(t)$ by $\pi_i(t)$, e.g., based on (3.4)-(3.6). If $S_i(t)$ were known, a bivariate Brier score would simply be the mean squared error, $\mathrm{MSE}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[S_i(t) - \pi_i(t)\right]^2$. With unknown $S_i(t)$, we estimate it with our observed data, and in the presence of censoring. Let $G_i(t) = \Pr(C_i > t) > 0$ be the survival function of the censoring distribution for the ith individual. We propose a bivariate Brier score for assessing $\pi_i(t)$ as follows:

$$
\begin{aligned}
\mathrm{BBS}(t) = &\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ \delta_{i1} = 1,\ Y_{i1} \le Y_{i2}\right\}}{G_i(Y_{i1})} \\
&+ \frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ Y_{i2} \le t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \le Y_{i2}\right\}}{G_i(Y_{i2})} \\
&+ \frac{\left[1 - \pi_i(t)\right]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right\}}{G_i(t)}.
\end{aligned} \tag{5.1}
$$

If $G_i(t)$ is known, the expectation of the IPCW-approximated bivariate Brier score is equal to $\mathrm{MSE}(t)$ plus a constant that is free of $\pi_i(t)$, which represents the irreducible error incurred by approximating $S_i(t)$ in a data-driven fashion (Supplement C). As $G_i(t)$ is unknown in practice, we replace it by $\hat{G}_i(t)$, its Kaplan-Meier estimate.

5.2. *Bivariate Concordance Index.* We further propose a bivariate extension to Harrell's concordance (C) index to evaluate the discriminative ability of the predicted hazards in a semi-competing risks setting (Harrell Jr, Lee and Mark, 1996). For each individual, we transform the predicted hazards for each transition into standardized scores by computing their empirical cumulative distribution function ranks and then applying the inverse normal transformation. The resulting quantile scores are averaged to obtain a risk score for each subject. We then consider all pairs of individuals (i, j) such that at least one of them experienced a non-censoring event (i.e., $D_{1i} = 1$ or $D_{2i} = 1$, and similarly for subject j). A pair is *comparable* if the event time of subject i is earlier than that of subject j in either the non-terminal or terminal event ($Y_{1i} < Y_{1j}$ or $Y_{2i} < Y_{2j}$). A comparable pair is considered *concordant* if the subject with the earlier event time has a higher predicted risk score. The bivariate C-index is then computed as the proportion of concordant pairs among all comparable pairs.

6. Simulation Studies. We conducted a series of numerical experiments to study the utility of our method compared to existing approaches under a range of settings.

6.1. *Simulation Setup.* We simulated data from Equation (3.7), varying the sample size (n) and log-risk functions (h_g) across six settings. Specifically, we generated independent datasets with $n = 1,000$ or $5,000$. We then simulated individual frailties, γ_i, from a Gamma distribution with mean 1 and variance $\theta = 2$, corresponding to a strong dependence between event times, as in our real data. To generate the risk functions, $h_g(\boldsymbol{X}_i); g \in \{1, 2, 3\}$, we considered three classes of covariate functions. In all cases, we generated twelve covariates, as in our motivating data, from a multivariate Normal distribution with a zero mean vector and a compound symmetric covariance matrix with diagonal elements (variances) equal to one and off-diagonal elements (covariances) equal to 0.2. That is,

$$\boldsymbol{X}_i \sim \mathcal{N}_{12}\left(\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}1.0 & 0.2 & \cdots & 0.2\\0.2 & 1.0 & \cdots & 0.2\\\vdots & \vdots & \ddots & \vdots\\0.2 & 0.2 & \ldots & 1.0\end{bmatrix}\right).$$

Linear Log-Risk Functions. We first considered linear $h_g(\boldsymbol{X}_i), g \in \{1, 2, 3\}$ so the requirements for the classical models were satisfied, facilitating a fair comparison. We let

$$h_g(\boldsymbol{X}_i) = X_{i,1}\beta_{1,g} + \cdots + X_{i,12}\beta_{12,g},$$

with $(\beta_{1,g}, \ldots, \beta_{12,g}) = (-0.5, -0.5, -0.5, -0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, 0)$, so $X_1, \ldots, X_9$ had true signals. We further generated the censoring times, C_i, to be covariate-dependent, coming from an exponential distribution with hazard

$$\lambda_c(\boldsymbol{X}_i) = \mu_C \times \exp\{X_{i,1}\alpha_1 + \cdots + X_{i,12}\alpha_{12}\},$$

where $(\alpha_1, \ldots, \alpha_{12}) = (0, 0, 0, 0, 0, 0, 0, -0.5, 0.5, -0.5, 0.5, -0.5)$, so $X_8, \ldots, X_{12}$ were related to the censoring time, with X_8 and X_9 related to both the survival and censoring times. We selected μ_C to achieve an approximate censoring rate of 25%. We chose the exponential family as existing methods are based on the Weibull-Cox or AFT models, allowing those methods to be correctly specified, providing a fair basis for evaluating our method.

Moderate Nonlinear Log-Risk Functions. We then simulated h_g with moderate nonlinearity by introducing first-order interaction and polynomial terms. Specifically, we let

$$\begin{aligned}
h_1(\boldsymbol{X}) &= -0.5X_1 - 0.5X_2 - 0.5X_3 - 0.5X_4 + 0.5X_5 + 0.5X_6 + 0.5X_7 + 0.5X_8\\
&\quad + 0.5X_9 + 0.03X_1X_2 - 0.02X_3X_4 + 0.02X_5X_6 + 0.005X_1^2 + 0.005X_2^2\\
&\quad + 0.005X_3^2 + 0.005X_4^2 - 0.005X_5^2 - 0.005X_6^2 - 0.005X_7^2 - 0.005X_8^2 - 0.005X_9^2\\
h_2(\boldsymbol{X}) &= -0.5X_1 - 0.5X_2 - 0.5X_3 - 0.5X_4 + 0.5X_5 + 0.5X_6 + 0.5X_7 + 0.5X_8\\
&\quad + 0.5X_9 + 0.0225X_1X_2 - 0.015X_3X_4 + 0.015X_5X_6 + 0.005X_1^2 + 0.005X_2^2\\
&\quad + 0.005X_3^2 + 0.005X_4^2 - 0.005X_5^2 - 0.005X_6^2 - 0.005X_7^2 - 0.005X_8^2 - 0.005X_9^2\\
h_3(\boldsymbol{X}) &= -0.5X_1 - 0.5X_2 - 0.5X_3 - 0.5X_4 + 0.5X_5 + 0.5X_6 + 0.5X_7 + 0.5X_8\\
&\quad + 0.5X_9 + 0.03X_1X_2 - 0.02X_3X_4 + 0.02X_5X_6 + 0.0025X_1^2 + 0.0025X_2^2 + 0.0025X_3^2\\
&\quad + 0.0025X_4^2 - 0.0025X_5^2 - 0.0025X_6^2 - 0.0025X_7^2 - 0.0025X_8^2 - 0.0025X_9^2,
\end{aligned}$$

with all other simulation settings unchanged. This setting reflects the degree of smooth, low-order nonlinearity that can still typically be captured by existing methods.

Nonlinear Log-Risk Functions. In a third scenario, we specified nonlinear h_g, with

$$\begin{aligned} h_g(\boldsymbol{X}_i) &= \beta_{1,g}\exp(X_{i,1} - X_{i,2}) - \beta_{2,g}\log\{(X_{i,3} + X_{i,4})^2\} \\ &+ \beta_{3,g}\sin(X_{i,5}X_{i,6}) - \beta_{4,g}(X_{i,7} - X_{i,8} + X_{i,9})^2, \\ \boldsymbol{\beta}_g &= (\beta_{1,g}, \ldots, \beta_{4,g}) = (= -0.5, -0.5, 0.5, 0.5) \end{aligned}$$

to highlight the utility of our method. Across all simulation settings, we generated baseline hazard functions, λ_{01}, λ_{02}, and λ_{03}, from exponential distributions so that $\lambda_{01} = \lambda_{03} = 2$, and $\lambda_{02} = 3$. For each parameter configuration, we generated 500 independent datasets.

6.2. *Methods.* We compared our method to six existing approaches for semi-competing risks and one for competing risks. Among the semi-competing methods, five specify (semi-) parametric models with linear log-risk functions: Xu, Kalbfleisch and Tai (2010), Lee et al. (2015), Lee, Rondeau and Haneuse (2017), Gorfine et al. (2021), and Kats and Gorfine (2022), and the sixth is a spline-based shared-frailty illness-death model with cubic B-spline bases (Lee, Gilsanz and Haneuse, 2021). To our knowledge, there are no tree-based methods that fit our setup, so we could not draw comparisons in terms of estimated risk functions, baseline hazards, or measures of dependence. However, our measures of predictive accuracy facilitate a fair comparison to competing risks methods, so we further compared to a random survival forest for multi-state outcomes (Ishwaran, Kogalur and Kogalur, 2022).

Each method was implemented following the authors' original specifications, with default or recommended settings. For the two Bayesian methods, we adopted relatively flat priors for the six Weibull baseline hazard parameters and tuned the prior mean and standard deviation of the frailty variance over a 10×10 grid (0.1-1.0) to maximize predictive performance. This ensured a fair comparison by allowing each method to operate under optimized configurations. For our approach, we set the number of layers to 3 and the dropout rate to 0.3, and we tuned the number of hidden nodes per layer (16-1024) and the learning rate (0.0001-0.05) using a grid search, selecting the configuration with the best predictive accuracy. We evaluated each method's performance in terms of estimating the mean (SD) frailty variance (θ) and baseline hazards (λ_{0g}), the mean integrated squared error (MISE) of the estimated h_g:

$$\text{MISE}_g = \frac{1}{n}\sum_{i=1}^{n}\left[h_g(\boldsymbol{X}_i) - \hat{h}_g(\boldsymbol{X}_i)\right]^2, \quad g = 1, 2, 3,$$

and our two proposed measures of predictive accuracy, the integrated bivariate Brier score (iBBS; up to $t = 1$ year) and the bivariate concordance index.

6.3. *Results.* Tables 1 and 2 summarize the results of this simulation study. Table 1 compares results for estimating the various semi-competing risk model components, namely the frailty variance (θ), baseline hazards (λ_{0g}), and log-risk functions (h_g), averaged over 500 independent replicates in each simulation setting. In terms of estimating θ, our proposed method had the lowest bias in three of the six settings, specifically in the two settings where the true risk functions were nonlinear and in one of the moderate nonlinear settings. Among the three settings where our proposed method did not have the lowest bias, the results were comparable to Gorfine et al. (2021), which had the best performance. Similiarly, in estimating the λ_{0g}, our proposed method had the lowest bias the nonlinear cases and comparable bias to Xu, Kalbfleisch and Tai (2010) in the remaining cases, despite being a nonparametric approximation. Lastly, our proposed method also performed comparably to the better performing methods in terms of the MISE for the predicted h_g, across all state transitions, when

the true underlying function of the predictors was linear or moderately nonlinear. In the nonlinear settings, our approach has the lowest MISEs, suggesting that our method outperforms (semi-)parametric approaches when the functional form of the predictors is truly nonlinear.

Table 2 summarizes the predictive accuracy of the methods. We calculated the integrated bivariate Brier score at one year over a sequence of 100 evenly spaced time points and compared the results of our method with the existing methods. As shown, the proposed method has the lowest integrated bivariate Brier score (lower = better) across all but one setting and the highest bivariate C-index (higher = better) across all settings, when compared to the other methods for semi-competing risks. We do note that the tree-based competing risks approach had better performance than our proposed approach in most settings, with both non-parametric approaches having better performance than the (semi-) parametric methods.

TABLE 1

Average (SD) estimated frailty variance (θ) *and baseline hazards* (λ_{0g})*, and average (SD) mean integrated squared errors for each log-risk function* $(h_g(\boldsymbol{X}_i); g = 1, 2, 3)$ *across six settings and 500 replicates per setting, varying the sample size* (n) *and log-risk functions* (h_g)*. All results are based on* $\theta = 2$ *and a* 25% *censoring rate.*

Simulation Settings		**Methods**						
n	h_g	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Lee (2021)	Proposed
Frailty Variance: θ								
1,000	Linear	1.45 (0.18)	0.79 (0.06)	0.57 (0.26)	**2.04 (0.57)**	1.40 (0.19)	0.42 (0.12)	2.06 (0.03)
5,000	Linear	1.44 (0.08)	0.78 (0.06)	0.57 (0.25)	**1.97 (0.21)**	1.48 (0.08)	0.41 (0.12)	2.08 (0.01)
1,000	Moderate	1.47 (0.18)	0.76 (0.24)	0.57 (0.25)	2.24 (0.48)	1.71 (0.09)	0.49 (0.13)	**2.18 (0.03)**
5,000	Moderate	1.43 (0.08)	0.72 (0.22)	0.56 (0.25)	**1.96 (0.20)**	1.73 (0.02)	0.47 (0.12)	2.13 (0.01)
1,000	Nonlinear	0.83 (0.26)	0.73 (0.28)	0.55 (0.27)	3.12 (0.62)	4.93 (5.79)	0.66 (0.22)	**2.14 (0.11)**
5,000	Nonlinear	0.81 (0.11)	0.73 (0.26)	0.56 (0.26)	2.83 (0.26)	4.52 (3.96)	0.59 (0.19)	**2.11 (0.01)**
Baseline Hazards: $\lambda_{0g}(t)$								
1,000	Linear	**0.36 (0.53)**	1.62 (1.59)	7.37 (3.50)	2.37 (0.10)	4.20 (2.45)	3.52 (2.03)	0.57 (0.58)
5,000	Linear	**0.32 (0.44)**	1.62 (1.59)	7.39 (3.49)	2.62 (2.09)	4.24 (3.17)	3.57 (2.02)	0.47 (0.39)
1,000	Moderate	**0.35 (0.51)**	1.54 (1.54)	7.36 (3.48)	2.68 (2.05)	4.26 (2.32)	3.32 (5.43)	0.62 (0.23)
5,000	Moderate	**0.32 (0.43)**	1.55 (1.55)	7.38 (3.52)	2.62 (2.09)	4.22 (2.35)	3.46 (1.98)	0.47 (0.19)
1,000	Nonlinear	1.72 (1.31)	1.94 (1.41)	8.79 (4.07)	2.37 (0.09)	4.21 (2.46)	3.51 (2.09)	**0.22 (0.46)**
5,000	Nonlinear	1.70 (1.20)	1.91 (1.36)	8.82 (4.03)	3.30 (1.50)	4.25 (2.44)	3.68 (2.10)	**0.12 (0.08)**
First Transition: $h_1(\boldsymbol{X}_i)$								
1,000	Linear	**0.10 (0.04)**	0.58 (0.10)	0.26 (0.09)	0.56 (0.09)	5.70 (0.65)	0.49 (0.13)	0.33 (0.11)
5,000	Linear	**0.04 (0.02)**	0.47 (0.06)	0.16 (0.03)	0.54 (0.04)	5.38 (0.49)	0.49 (0.11)	0.22 (0.04)
1,000	Moderate	5.39 (0.55)	3.35 (0.38)	0.26 (0.09)	3.03 (0.24)	**0.22 (0.10)**	3.73 (0.66)	2.07 (0.17)
5,000	Moderate	5.26 (0.23)	3.45 (0.25)	0.16 (0.04)	2.98 (0.11)	**0.06 (0.02)**	3.53 (0.39)	1.22 (0.04)
1,000	Nonlinear	4.31 (0.41)	4.54 (0.41)	4.43 (0.40)	4.56 (0.41)	7.99 (1.14)	4.46 (0.41)	**3.82 (0.59)**
5,000	Nonlinear	4.31 (0.18)	4.49 (0.18)	4.33 (0.17)	4.58 (0.18)	8.06 (0.66)	4.49 (0.21)	**2.88 (0.20)**
Second Transition: $h_2(\boldsymbol{X}_i)$								
1,000	Linear	**0.08 (0.03)**	0.45 (0.08)	0.13 (0.05)	0.54 (0.08)	5.59 (0.50)	0.35 (0.10)	0.29 (0.11)
5,000	Linear	**0.03 (0.01)**	0.36 (0.05)	0.05 (0.02)	0.54 (0.03)	5.36 (0.46)	0.35 (0.09)	0.17 (0.04)
1,000	Moderate	5.52 (0.52)	3.63 (0.38)	**0.13 (0.05)**	3.01 (0.22)	**0.13 (0.05)**	3.09 (0.78)	2.24 (0.20)
5,000	Moderate	5.40 (0.21)	3.71 (0.24)	**0.05 (0.02)**	2.97 (0.10)	**0.05 (0.02)**	3.84 (0.38)	1.42 (0.06)
1,000	Nonlinear	4.28 (0.40)	4.47 (0.40)	4.25 (0.37)	4.55 (0.40)	7.75 (0.98)	4.39 (0.38)	**3.77 (0.57)**
5,000	Nonlinear	4.29 (0.17)	4.45 (0.18)	4.20 (0.16)	4.58 (0.18)	7.78 (0.45)	4.42 (0.19)	**2.95 (0.23)**
Third Transition: $h_3(\boldsymbol{X}_i)$								
1,000	Linear	**0.10 (0.04)**	0.58 (0.10)	0.29 (0.12)	0.56 (0.09)	5.70 (0.65)	0.49 (0.13)	0.40 (0.14)
5,000	Linear	**0.04 (0.02)**	0.47 (0.06)	0.13 (0.04)	0.54 (0.04)	5.38 (0.50)	0.49 (0.11)	0.25 (0.06)
1,000	Moderate	5.40 (0.55)	3.35 (0.38)	0.26 (0.09)	3.03 (0.25)	**0.22 (0.09)**	3.73 (0.66)	1.96 (0.15)
5,000	Moderate	5.26 (0.23)	3.45 (0.25)	0.16 (0.04)	2.98 (0.11)	**0.06 (0.02)**	3.53 (0.39)	1.25 (0.06)
1,000	Nonlinear	4.31 (0.41)	4.54 (0.41)	4.64 (0.45)	4.56 (0.41)	7.99 (1.14)	4.46 (0.41)	**4.05 (0.56)**
5,000	Nonlinear	4.31 (0.18)	4.49 (0.18)	4.45 (0.19)	4.58 (0.18)	8.06 (0.66)	4.49 (0.21)	**3.16 (0.25)**

6.4. *Sensitivity Analyses.* We further evaluated our proposed method and bivariate Brier score in a series of sensitivity analyses. We briefly summarize these here. (1) We examined the performance of our proposed bivariate Brier score and compared the results from the model fit to a calculation that utilized the true model parameters across four settings. The results from the model fit were on par with those calculated using the true model parameters, giving an approximate lower bound for the bivariate Brier score. (2) We studied the robustness of our method to the assumed gamma distribution of the latent frailties. Our approach

TABLE 2

Integrated bivariate Brier score and bivariate concordance index over six settings and 500 replicates per setting, varying the sample size (n) and log-risk functions (h_g). All results are based on $\theta = 2$ and 25% censoring.

Simulation Settings		Methods							
n	h_g	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Lee (2021)	Ishwaran (2022)	Proposed
Integrated Bivariate Brier Score (iBBS)									
1,000	Linear	0.23 (0.009)	0.21 (0.010)	0.23 (0.010)	0.21 (0.05)	0.51 (0.02)	0.32 (0.04)	**0.13 (0.005)**	0.16 (0.006)
5,000	Linear	0.16 (0.003)	0.22 (0.007)	0.20 (0.006)	0.26 (0.01)	0.51 (0.02)	0.34 (0.04)	**0.13 (0.002)**	0.16 (0.003)
1,000	Moderate	0.16 (0.01)	0.16 (0.01)	0.58 (0.05)	0.27 (0.03)	0.53 (0.02)	0.24 (0.03)	**0.13 (0.004)**	0.17 (0.005)
5,000	Moderate	0.16 (0.003)	0.16 (0.003)	0.54 (0.03)	0.26 (0.01)	0.53 (0.02)	0.32 (0.04)	0.13 (0.002)	**0.12 (0.004)**
1,000	Nonlinear	0.24 (0.010)	0.18 (0.007)	0.21 (0.008)	0.33 (0.03)	0.55 (0.11)	0.35 (0.05)	**0.14 (0.004)**	0.16 (0.006)
5,000	Nonlinear	0.18 (0.003)	0.18 (0.003)	0.19 (0.003)	0.32 (0.02)	0.52 (0.09)	0.38 (0.05)	0.13 (0.002)	**0.10 (0.005)**
Bivariate Concordance (C) Index									
1,000	Linear	0.63 (0.01)	0.63 (0.02)	0.51 (0.02)	0.61 (0.06)	0.41 (0.02)	0.61 (0.02)	**0.77 (0.01)**	0.64 (0.01)
5,000	Linear	0.63 (0.01)	0.62 (0.01)	0.51 (0.02)	0.54 (0.01)	0.42 (0.02)	0.61 (0.01)	**0.77 (0.01)**	0.64 (0.01)
1,000	Moderate	0.59 (0.02)	0.63 (0.01)	0.62 (0.01)	0.54 (0.02)	0.58 (0.02)	0.54 (0.02)	**0.78 (0.01)**	0.63 (0.01)
5,000	Moderate	0.59 (0.01)	0.63 (0.01)	0.63 (0.01)	0.54 (0.01)	0.58 (0.01)	0.59 (0.01)	**0.77 (0.01)**	0.66 (0.01)
1,000	Nonlinear	0.59 (0.02)	0.59 (0.02)	0.53 (0.02)	0.51 (0.02)	0.46 (0.03)	0.55 (0.02)	**0.79 (0.01)**	0.69 (0.01)
5,000	Nonlinear	0.63 (0.01)	0.60 (0.01)	0.53 (0.01)	0.51 (0.01)	0.46 (0.02)	0.55 (0.01)	**0.80 (0.01)**	0.68 (0.01)

remained robust in terms of prediction accuracy and baseline hazard estimation, with bias appearing primarily in the log-risk functions when the frailty variance was large or the true risk function was highly nonlinear. (3) We performed a benchmarking study to determine the computational cost of our method and the existing methods. All runtime comparisons reported here were conducted using CPU-based implementations. Under these settings, the wall time for our proposed approach scaled approximately linearly with n and sub-linearly with p, with training averaging 20 seconds on a standard MacBook Pro for the largest setting. This was comparable to the standard existing methods and faster than the more contemporary methods. Further, while GPU acceleration could further reduce runtime in larger-scale applications, it was not required for the analyses presented here. (4) We compared our approach to the existing (semi-) parametric methods across additional settings, where we further varied the population frailty variance ($\theta = 0.5$ or 2) and μ_C to achieve approximate censoring rates of 25% or 50%. In general, our proposed approach performed comparably to the existing approaches on all metrics in the linear settings and outperformed in the nonlinear settings. Full details and results can be found in Supplement D.

7. Analysis of Boston Lung Cancer Study (BLCS). Among the 19,497 participants in the BLCS, 7,755 met the initial eligibility criterion of a confirmed lung cancer diagnosis. Participants were excluded if enrolled with other primary cancers (e.g., esophageal), lacked a cancer diagnosis, or were negative controls (e.g., spouses or friends). Of these, 7,697 (99.3%) had complete temporal information required for defining semi-competing outcomes, including dates of diagnosis, progression or recurrence, death, and last follow-up. We excluded 58 patients (0.7%) missing diagnosis dates, 56 (0.7%) with inadequate follow-up (diagnosis within six months of study end), 207 (2.7%) with small-cell lung cancer, six (0.08%) with carcinoma *in situ*, and 25 (0.3%) with identical diagnosis and event dates. The final analytic cohort included $n = 7,403$ (95.5%) patients with NSCLC diagnosed between June 1983 and February 2023. Among them, 2,443 (33.0%) experienced progression, 1,570 (21.2%) had progression followed by death, and 3,636 (49.1%) died prior to progression (Table 3).

Detailed information on patient demographics, smoking history, and physiologic measurements were collected through questionnaire when the patient was recruited to the Boston Lung Cancer Study, at their time of diagnosis. Genetic mutations were also collected. We considered eleven demographic, clinical, and genetic risk factors for this analysis. Potential demographic predictors included patient age at diagnosis (years), sex assigned at birth, self-identified race, and ethnicity. Smoking status and pack-years of smoking were also included.

TABLE 3
Semi-competing event rates among $n = 7,403$ patients in our analytic sample.

Progression Observed / Death Observed	Yes	No
Yes	1,570 (21.21%)	873 (11.79%)
No	3,636 (49.12%)	1,324 (17.88%)

Relevant clinical predictors included cancer stage at diagnosis, initial treatment, indications of chronic obstructive pulmonary disease (COPD) or asthma, and oncogenic (somatic driver) mutation status (EGFR or KRAS). Table 4 reports summary statistics for these risk factors in our study sample. As shown, median (interquartile range; IQR) age at diagnosis was 67 (59, 74) years, with 3,966 (54%) patients being female, 6,834 (92%) being White, and 6,410 (87%) being non-Hispanic. Clinically relevant features are as follows. The majority of patients had a history of smoking (6,259; 85%), with a median (IQR) of 36 (11, 57) pack-years of smoking. Further, 1,553 patients (21%) were tested using the SNaPshot assay for the presence of genetic variants. The results of this testing revealed that 405 (5.5%) patients were positive for at least one KRAS variant and 298 (4.0%) patients were positive for at least one EGFR variant. COPD was present in 2,284 (55%) patients and 410 (7.7%) patients had asthma. Lastly, 4,444 (60%) patients initially underwent surgery, while 1,851 (25%) patients initially received chemotherapy, 366 (4.9%) received radiation, and 742 (10%) received another form of treatment (Table 4). The distributions of these characteristics are similar to a recent study utilizing patient data from Massachusetts General Hospital, which draws comparisons to the BLCS cohort (Yuan et al., 2021).

7.1. *Predictive Modeling.* After one-hot encoding, our design matrix consisted of 25 features used in predictive modeling. We applied our proposed method together with those of Xu, Kalbfleisch and Tai (2010), Lee et al. (2015), Lee, Rondeau and Haneuse (2017), Lee, Gilsanz and Haneuse (2021), and Ishwaran, Kogalur and Kogalur (2022) to estimate hazards of progression, death, and death following progression using the predictors in Table 4. The methods of Gorfine et al. (2021) and Kats and Gorfine (2022) failed to converge and were excluded. We evaluated performance via five-fold cross-validation, training on 80% of samples and validating on 20%. We optimized our hyperparameters, including the number of nodes per layer, learning rate, dropout, and regularization, by grid search. Our optimal subnetworks contained three hidden layers with 1024, 64, and 32 nodes, respectively, a dropout rate of 0.1, and a learning rate of 0.0001. We assessed model accuracy using the integrated bivariate Brier score at 100 time points up to five years post-diagnosis and bivariate C-index, averaged across folds. Table 5 shows that the tree-based approach had the best predictive accuracy on both metrics, with the approach of Xu, Kalbfleisch and Tai (2010) having similar performance in terms of the bivariate Brier score and our approach having lower, but comparable performance. This is expected, as deep learning may not always be optimal for purely tabular data, and tree-based learners may be preferable in smaller samples (Grinsztajn, Oyallon and Varoquaux, 2022). However, for higher-dimensional or more heterogeneous covariate structures, neural networks can serve as a potential scalable, data-adaptive choice.

We then applied the proposed NEM-SCR algorithm to estimate the frailty variance, θ, obtaining a value of 2.09. For context, this is approximately equal to a Kendall's τ value of 0.511 (Austin, 2017). To quantify the associated uncertainty, we computed a bootstrap standard error of 0.04 based on 50 resamplings of the data with replacement. This nonzero estimate of frailty suggests the presence of moderate subject-level dependence across the three transitions. Figure 3 displays the average estimated cumulative baseline hazard functions along with 95% bootstrap confidence intervals constructed from these 50 replicates. As

TABLE 4

Characteristics of the $n = 7,403$ patients with non-small cell lung cancer diagnosed between June 1983 and February 2023 in our analytic sample derived from the Boston Lung Cancer Study cohort. Summary statistics are reported as $n(\%)$ for categorical predictors and median (interquartile range) for continuous covariates.

Characteristic	**Total N = 7,403**[1]	**Cancer Stage: Early N = 4,700**	**Cancer Stage: Late N = 2,342**	**Cancer Stage: Unknown N = 361**
Age at Diagnosis (yrs.)	67 (59, 74)	68 (61, 75)	64 (56, 72)	66 (58, 73)
Unknown	220	212	7	1
Sex				
Female	3,966 (54%)	2,603 (55%)	1,188 (51%)	175 (48%)
Male	3,431 (46%)	2,093 (45%)	1,152 (49%)	186 (52%)
Unknown	6	4	2	0
Race				
White/Caucasian	6,834 (92%)	4,349 (93%)	2,149 (92%)	336 (93%)
Black/African American	126 (1.7%)	83 (1.7%)	40 (1.7%)	3 (0.8%)
Asian	140 (1.9%)	69 (1.5%)	67 (2.9%)	4 (1.1%)
Other	98 (1.3%)	60 (1.3%)	35 (1.5%)	3 (0.8%)
Unknown	205 (2.7%)	139 (3.0%)	51 (2.2%)	15 (4.1%)
Ethnicity				
Non-Hispanic	6,410 (87%)	3,990 (85%)	2,112 (90%)	308 (85%)
Hispanic	87 (1.2%)	57 (1.2%)	28 (1.2%)	2 (0.6%)
Unknown	906 (13%)	653 (14%)	202 (8.6%)	51 (14%)
Smoking Status				
Smoker	6,259 (85%)	4,007 (85%)	1,917 (82%)	335 (93%)
Non-Smoker	1,009 (14%)	592 (13%)	402 (17%)	15 (4.2%)
Unknown	135 (1.8%)	101 (2.1%)	23 (1.0%)	11 (3.0%)
Pack-Years of Smoking	36 (11, 57)	37 (12, 58)	32 (8, 52)	49 (30, 76)
Unknown	958	818	99	41
Initial Treatment				
Surgery	4,444 (60%)	3,994 (85%)	378 (16%)	72 (20%)
Chemotherapy	1,851 (25%)	365 (7.8%)	1,473 (62%)	13 (3.6%)
Radiation	366 (4.9%)	194 (4.1%)	163 (6.9%)	9 (2.4%)
Other/Unknown	742 (10%)	147 (3.1%)	328 (14%)	267 (74%)
EGFR Status				
Variant Negative	1,255 (17%)	737 (16%)	498 (21%)	20 (5.5%)
Variant Positive	298 (4.0%)	158 (3.4%)	140 (6.0%)	0 (0%)
Not Tested	5,850 (79%)	3,805 (81%)	1,704 (73%)	341 (94%)
KRAS Status				
Variant Negative	1,148 (16%)	630 (13%)	500 (21%)	18 (5.0%)
Variant Positive	405 (5.5%)	265 (5.6%)	138 (5.9%)	2 (0.6%)
Not Tested	5,850 (79%)	3,805 (81%)	1,704 (73%)	341 (94%)
COPD[2]	2,284 (55%)	1,662 (61%)	505 (40%)	117 (54%)
Asthma	410 (7.7%)	254 (7.6%)	136 (7.9%)	20 (7.1%)

[1]Median (IQR); n (%) [2]COPD: Chronic Obstructive Pulmonary Disease;

TABLE 5

Integrated bivariate Brier score and bivariate C-index for each method.

Method	**Integrated Bivariate Brier Score (iBBS)**	**Bivariate C-Index**
Xu (2010)	**0.18 (0.17-0.18)**	0.64 (0.61-0.67)
Lee (2015)	0.53 (0.52-0.53)	0.58 (0.54-0.61)
Lee (2017)	0.53 (0.52-0.54)	0.57 (0.53-0.60)
Spline-Based Method	0.25 (0.22-0.28)	0.63 (0.60-0.66)
Tree-Based Method	**0.18 (0.17-0.19)**	**0.69 (0.66-0.72)**
Proposed Method	0.30 (0.29-0.31)	0.64 (0.60-0.68)

shown, the baseline hazards are highest for death, with high variability in these estimates, comparable to our simulations with similar frailty variance values and censoring rates.

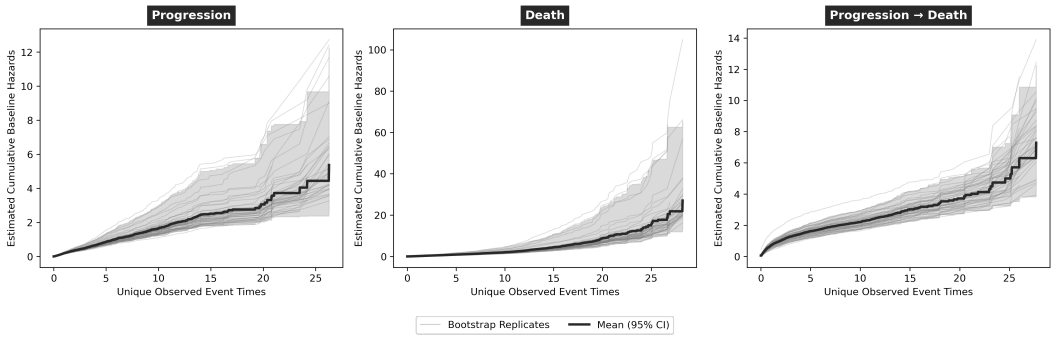

Fig 3: Average estimated cumulative baseline hazards (solid dark lines) and 95% bootstrap confidence intervals (gray bands) based on 50 bootstrap samples (light gray lines).

We compare existing approaches in modeling risk factor effects across the three state transitions (progression, death, and death following progression) in Table E.1 in the Supplementary Material. The estimated log hazard ratios (log-HR) often differ in both sign and magnitude, revealing inconsistencies among methods that assume linear risk functions. For example, age at diagnosis is a significant risk factor for all transitions under Xu, Kalbfleisch and Tai (2010) and Lee et al. (2015), but not under Lee, Rondeau and Haneuse (2017). Moreover, while Xu, Kalbfleisch and Tai (2010) report a small positive effect on progression (log-HR = 0.01; SE = 0.002), Lee et al. (2015) find a small negative effect (log-HR = -0.04; SE = 0.002). For sex, Xu, Kalbfleisch and Tai (2010) estimate lower risks for females across all transitions (e.g., log-HR = -0.19 for progression), while Lee et al. (2015) detect significance only for death after progression (log-HR = -0.37), and Lee, Rondeau and Haneuse (2017) instead report a higher death risk for females (log-HR = 0.79). For smoking, all three methods agree on increased risk of death after progression among current versus never smokers, but diverge on other transitions: Xu, Kalbfleisch and Tai (2010) find increased risks for progression and death from diagnosis, Lee et al. (2015) report a protective effect for progression, and Lee, Rondeau and Haneuse (2017) a protective effect for death from diagnosis. Similar discrepancies appear for other predictors, underscoring the instability of linear specifications.

Figure 4 depicts the log-risk (h_g) functions for the predicted effect of patient age at diagnosis on each state transition, take over a sequence of potential ages (40 to 75 years) and stratified by sex (male versus female) and smoking status (smoker versus non-smoker). All other covariates were fixed to be at their sample means or modes for illustration. As shown, there is a slight, increasing relationship between age and all three state transitions, but particularly in the transition from progression to death. Further, these relationships differ by sex, with males having a higher risk of death and death following progression. Smoking status appears to have a stronger effect on the risk of death from diagnosis, with the separation in risk between male smokers versus non-smokers and male versus female patients suggesting potential interaction effects for this state transition.

8. Discussion. We developed a neural expectation-maximization algorithm for predicting semi-competing risks, where a non-terminal event (e.g., disease progression) modifies the risk of death. Unlike existing machine learning models that treat events as independent, our framework jointly captures the dependence between progression and mortality, yielding accurate estimates of transition-specific hazards, baseline hazards, covariate effects, and the degree of dependence between events, even under complex risk structures. As a first deep learning approach for semi-competing risks, our goal in this work was to present an honest account of its advantages and limitations. In simulation, our proposed method outperformed

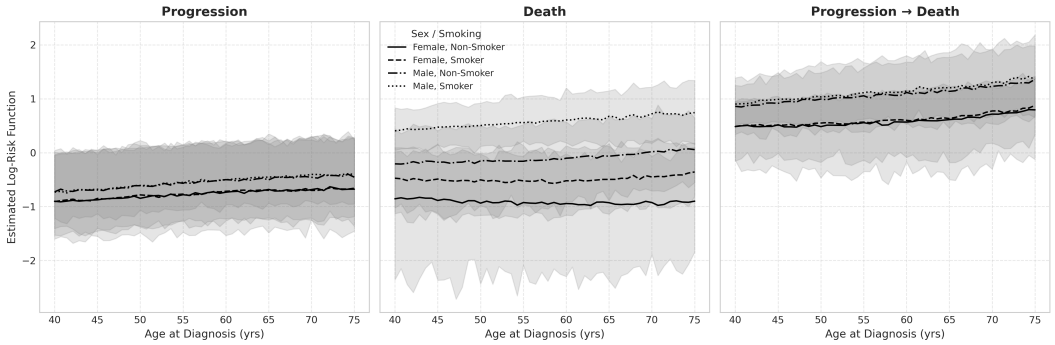

Fig 4: Nonparametric log-risk functions for age at diagnosis on each state transition, stratified by sex and smoking status (line types), and 95% bootstrap confidence intervals (gray bands).

the current semi-competing methods in terms of predictive accuracy and in estimating key model components, including the frailty variance, risk functions, and baseline hazards when the underlying covariate effects were nonlinear. This likely reflects improved recovery of complex data-generating mechanisms via the nonparametric estimation of these quantities.

However, we acknowledge that deep learning may not always be optimal for purely tabular data, especially with modest sample sizes (Grinsztajn, Oyallon and Varoquaux, 2022). This is evident in our simulations, where random survival forests achieved better predictive accuracy, and in our real data example, where existing semi-competing risk methods also performed competitively. Our contribution complements predictive benchmarking alone. Our framework accounts for joint estimation of nonparametric baseline hazards and nonlinear risk functions within an EM structure, providing interpretable quantities, such as baseline hazards and frailty variance, not typically available from tree-based models. While tree-based learners such as boosting and random forests can flexibly capture nonlinear covariate effects, they do not directly model dependence between transitions in semi-competing risks settings. The proposed approach instead targets joint estimation of transition-specific risks and their dependence through a shared frailty illness-death framework, while retaining flexibility in modeling covariate effects. Further, while our preliminary analysis focused on a small set of structured covariates, the approach naturally extends to multi-modal data (e.g., radiomic or genomic features), where deep architectures are most effective and where our future motivation lies. Deep neural networks circumvent the *curse of dimensionality* in nonparametric settings (Bauer and Kohler, 2019; Poggio et al., 2017) by projecting the data into a much lower relevant representational space (Abrol et al., 2021; Goodfellow, Bengio and Courville, 2016). As our EM framework remains correct for any flexible estimator of h_g, neural networks can serve as a potential scalable, data-adaptive choice.

Applied to the Boston Lung Cancer Study, NEM-SCR uncovered interactions among age, sex, and smoking status, showing that male smokers had a higher associated mortality risk, consistent with prior reports (Guo, Tosun and Horn, 2009; Tseng et al., 2022), but obtained through explicit joint modeling of progression and death. The method had comparable predictive accuracy relative to standard regression models, while avoiding the information loss inherent in composite endpoints such as progression-free survival. Finally, we note that the estimated h_g functions represent risk scores on the log-hazard scale, interpretable only up to additive constants, and should be viewed as associational rather than causal effects. Meaningful interpretation further requires calibration assessment, proper variable scaling, and caution when outside a formal causal framework (Dumas and Stensrud, 2025).

In summary, as a first deep learning approach for semi-competing risks, the goal of this work was to to provide a transparent, side-by-side assessment showing that different meth-

ods have different strengths, as supported by both simulations and real data. Our proposed approach offers a flexible framework that can potentially improve the estimation of complex latent and structural components in semi-competing risks models while maintaining competitive or comparable predictive performance.

There are several open directions for future work. First, we assume a parametric frailty, which we saw was robust to misspecification in terms of overall predictive accuracy and in estimating the baseline hazards, but sensitive in terms of estimating the log-risk functions. An alternative may be a fully nonparametric approach by specifying a frailty with a finite, but unknown, number of mixture components (Gasperoni et al., 2020; Chee et al., 2021). Second, while we focus on time-independent covariates, future work may consider time-varying predictors, which would increase the utility of the method. We also consider a subset of structured features from the BLCS study. Future work will also consider multi-modal and high-dimensional predictors such as imaging and genetic data. Third, while our use of the bootstrap provides a practical means to quantify uncertainty in our estimates, we acknowledge that it remains an ad hoc solution without formal guarantees in deep learning settings. Developing rigorous inferential theory for neural network-based estimators, particularly in semiparametric or nonparametric models, remains an open and promising area for future research. Further, to our knowledge, there are no random forest methods for semi-competing risks that can estimate all model components targeted by our joint framework. As our proposed algorithm separates estimation of the baseline hazards and frailty variance from the log-risk functions, the N-step can, in principle, accommodate alternative nonlinear learners such as gradient boosting or random forests, while retaining the interpretable illness-death model structure. Such extensions require redesigning the optimization procedure and are therefore beyond the scope of the present work but represent a promising direction for future research. Lastly, while we focus on the joint distribution of the observed survival times for both event processes simultaneously, sometimes it can be of interest to study the marginal distribution of the non-terminal event (e.g., disease progression) while addressing the dependent censoring incurred by death. We will address these problems elsewhere.

Significance Statement. We propose a neural expectation-maximization framework for semi-competing risks that integrates classical survival modeling with deep learning to jointly predict disease progression and mortality. Motivated by the Boston Lung Cancer Study, the method captures complex dependencies among clinical and genetic factors, quantifies event dependence, and enables accurate, interpretable dynamic risk stratification for personalized treatment planning that is applicable to diverse diseases.

Acknowledgments. We thank our long-term collaborator, Dr. David C. Christiani, for providing the Boston Lung Cancer Study data and Jui Kothari for preparing these data. We thank Drs. Christiani and Xinan Wang for helpful discussions. We also thank the Area Editors, Drs. Jeffrey Morris and Tapabrata Maiti, and the referees for their insightful suggestions that helped substantially improve the quality and presentation of the manuscript.

Funding. Yi Li is supported by the National Institutes of Health grants R01CA249096. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no competing interests.

SUPPLEMENTARY MATERIAL

A. Illness-Death Model
Additional technical details related to the illness-death model and conditional likelihood.

B. NEM-SCR Algorithm Details
Additional technical details in support of the proposed NEM-SCR algorithm.

C. Bivariate Brier Score
Additional technical details in support of our proposed bivariate Brier score.

D. Additional Simulation Results
Additional sensitivity results in support of our main numerical experiments.

E. Additional Data Analysis Results
Classical semi-competing risk model estimates for the BLCS analysis.

F. Summary of Comparisons Between Semi-Competing Risk Models
Comparison of key modeling assumptions and methodological characteristics, including model structure, flexibility, computational features, and data requirements, between the proposed approach and classical semi-competing risk models.

REFERENCES

AASTHA, P. H. and LIU, Y. (2020). DeepCompete: A deep learning approach to competing risks in continuous time domain. In *AMIA Annual Symposium Proceedings* **2020** 177. American Medical Informatics Association.

ABROL, A., FU, Z., SALMAN, M., SILVA, R., DU, Y., PLIS, S. and CALHOUN, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications* **12** 1–17.

ALVARES, D., HANEUSE, S., LEE, C. and LEE, K. H. (2019). SemiCompRisks: an R package for the analysis of independent and cluster-correlated semi-competing risks data. *The R journal* **11** 376.

ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

ASHWORTH, A. B., SENAN, S., PALMA, D. A., RIQUET, M., AHN, Y. C., RICARDI, U., CONGEDO, M. T., GOMEZ, D. R., WRIGHT, G. M., MELLONI, G. et al. (2014). An individual patient data metaanalysis of outcomes and prognostic factors after treatment of oligometastatic non–small-cell lung cancer. *Clinical lung cancer* **15** 346–355.

AUSTIN, P. C. (2017). A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review* **85** 185–203.

BA, J. and CARUANA, R. (2014). Do deep nets really need to be deep? *Advances in neural information processing systems* **27**.

BADE, B. C. and CRUZ, C. S. D. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. *Clinics in chest medicine* **41** 1–24.

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* **47** 2261–2285.

BRIER, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* **78** 1–3.

CARREIRA-PERPINAN, M. A. and HINTON, G. (2005). On contrastive divergence learning. In *International workshop on artificial intelligence and statistics* 33–40. PMLR.

CHEE, C.-S., DO HA, I., SEO, B. and LEE, Y. (2021). Semiparametric estimation for nonparametric frailty models using nonparametric maximum likelihood approach. *Statistical methods in medical research* **30** 2485–2502.

CHEN, J., ZHANG, R., XU, J., HU, C. and MAO, Y. (2022). A neural expectation-maximization framework for noisy multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering* **35** 10992–11003.

CHRISTIANI, D. C. (2017). The Boston lung cancer survival cohort. http://grantome.com/grant/NIH/U01-CA209414-01A1. [Online; accessed November 19, 2025].

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34** 187–202.

DUMAS, E. and STENSRUD, M. J. (2025). How hazard ratios can mislead and why it matters in practice. *European Journal of Epidemiology* 1–7.

EISENHAUER, E. A., THERASSE, P., BOGAERTS, J., SCHWARTZ, L. H., SARGENT, D., FORD, R., DANCEY, J., ARBUCK, S., GWYTHER, S., MOONEY, M. et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer* **45** 228–247.

FARAGGI, D. and SIMON, R. (1995). A neural network model for survival data. *Statistics in Medicine* **14** 73–82.

FINE, J. P., JIANG, H. and CHAPPELL, R. (2001). On semi-competing risks data. *Biometrika* **88** 907–919.

GASPAR, L. E., MCNAMARA, E. J., GAY, E. G., PUTNAM, J. B., CRAWFORD, J., HERBST, R. S. and BONNER, J. A. (2012). Small-cell lung cancer: prognostic factors and changing treatment over 15 years. *Clinical lung cancer* **13** 115–122.

GASPERONI, F., IEVA, F., PAGANONI, A. M., JACKSON, C. H. and SHARPLES, L. (2020). Non-parametric frailty Cox models for hierarchical time-to-event data. *Biostatistics* **21** 531–544.

GOEL, A., RAIZADA, A., AGRAWAL, A., BANSAL, K., UNIYAL, S., PRASAD, P., YADAV, A., TYAGI, A. and RAUTELA, R. (2021). Correlates of in-hospital COVID-19 deaths: a competing risks survival time analysis of retrospective mortality data. *Disaster Medicine and Public Health Preparedness* 1–27.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep learning*. MIT press.

GORFINE, M., KERET, N., BEN ARIE, A., ZUCKER, D. and HSU, L. (2021). Marginalized frailty-based illness-death model: application to the UK-Biobank survival data. *Journal of the American Statistical Association* **116** 1155–1167.

GREFF, K., VAN STEENKISTE, S. and SCHMIDHUBER, J. (2017). Neural expectation maximization. *Advances in neural information processing systems* **30**.

GRINSZTAJN, A., OYALLON, E. and VAROQUAUX, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems* **35** 507–520.

GUO, N. L., TOSUN, K. and HORN, K. (2009). Impact and interactions between smoking and traditional prognostic factors in lung cancer progression. *Lung cancer* **66** 386–392.

HADDAD, R., CRUM, C., CHEN, Z., KRANE, J., POSNER, M., LI, Y. and BURK, R. (2008). HPV16 transmission between a couple with HPV-related head and neck cancer. *Oral Oncol* **44** 812–815.

HANEUSE, S. and LEE, K. H. (2016). Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes* **9** 322–331.

HAO, L., KIM, J., KWON, S. and HA, I. D. (2021). Deep learning-based survival analysis for high-dimensional survival data. *Mathematics* **9** 1244.

HARRELL JR, F. E., LEE, K. L. and MARK, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15** 361–387.

INAMURA, K. and ISHIKAWA, Y. (2010). Lung cancer progression and metastasis from the prognostic point of view. *Clinical & experimental metastasis* **27** 389–397.

ISHWARAN, H., KOGALUR, U. B. and KOGALUR, M. U. B. (2022). Package 'randomForestSRC'. *breast* **6**.

JAZIĆ, I., SCHRAG, D., SARGENT, D. J. and HANEUSE, S. (2016). Beyond composite endpoints analysis: semi-competing risks as an underutilized framework for cancer research. *JNCI: Journal of the National Cancer Institute* **108**.

JIANG, F. and HANEUSE, S. (2017). A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics* **44** 112–129.

JIANG, Z., ZHENG, Y., TAN, H., TANG, B. and ZHOU, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

JING, H. and SMOLA, A. J. (2017). Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* 515–524.

JOHANSEN, S. (1983). An extension of Cox's regression model. *International Statistical Review/Revue Internationale de Statistique* 165–174.

KATS, L. and GORFINE, M. (2022). An accelerated failure time regression model for illness–death data: A frailty approach. *Biometrics*.

KATZMAN, J. L., SHAHAM, U., CLONINGER, A., BATES, J., JIANG, T. and KLUGER, Y. (2016). Deep survival: A deep cox proportional hazards network. *STAT* **1050** 1–10.

KIM, S., ZENG, D., CHAMBLESS, L. and LI, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Biosciences* **4** 262–281.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint:1412.6980*.

KVAMME, H., BORGAN, Ø. and SCHEEL, I. (2019). Time-to-event prediction with neural networks and Cox regression. *arXiv preprint arXiv:1907.00825*.

LEE, C., GILSANZ, P. and HANEUSE, S. (2021). Fitting a shared frailty illness-death model to left-truncated semi-competing risks data to examine the impact of education level on incident dementia. *BMC medical research methodology* **21** 18.

LEE, K. H., RONDEAU, V. and HANEUSE, S. (2017). Accelerated failure time models for semi-competing risks data in the presence of complex censoring. *Biometrics* **73** 1401–1412.

LEE, C., YOON, J. and VAN DER SCHAAR, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering* **67** 122–133.

LEE, K. H., HANEUSE, S., SCHRAG, D. and DOMINICI, F. (2015). Bayesian semiparametric analysis of semicompeting risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society Series C: Applied Statistics* **64** 253–273.

LEE, C., ZAME, W. R., YOON, J. and VAN DER SCHAAR, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-second AAAI conference on artificial intelligence*.

LI, Y. and LIN, X. (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika* **87** 849–866.

LI, J., ZHANG, Y., BAKOYANNIS, G. and GAO, S. (2020). On shared gamma-frailty conditional Markov model for semicompeting risks data. *Statistics in Medicine* **39** 3042–3058.

LYNCH, T. J., BELL, D. W., SORDELLA, R., GURUBHAGAVATULA, S., OKIMOTO, R. A., BRANNIGAN, B. W., HARRIS, P. L., HASERLAT, S. M., SUPKO, J. G. and HALUSKA, F. G. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine* **350** 2129–2139.

ORSTAD, S., FLØTTEN, Ø., MADEBO, T., GULBRANDSEN, P., STRAND, R., LINDEMARK, F., FLUGE, S., TILSETH, R. H. and SCHAUFEL, M. A. (2023). "The challenge is the complexity"–A qualitative study about decision-making in advanced lung cancer treatment. *Lung Cancer* **183** 107312.

PAEZ, J. G., JANNE, P. A., LEE, J. C., TRACY, S., GREULICH, H., GABRIEL, S., HERMAN, P., KAYE, F. J., LINDEMAN, N., BOGGON, T. J. et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304** 1497–1500.

PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., DEVITO, Z., LIN, Z., DESMAISON, A., ANTIGA, L. and LERER, A. (2017). Automatic differentiation in pytorch.

POGGIO, T., MHASKAR, H., ROSASCO, L., MIRANDA, B. and LIAO, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing* **14** 503–519.

RANGANATH, R., PEROTTE, A., ELHADAD, N. and BLEI, D. (2016). Deep survival analysis. In *Machine Learning for Healthcare Conference* 101–114. PMLR.

REEDER, H. T., LU, J. and HANEUSE, S. (2022). Penalized estimation of frailty-based illness-death models for semi-competing risks. *arXiv preprint arXiv:2202.00618*.

SALERNO, S. and LI, Y. (2022). High-Dimensional Survival Analysis: Methods and Applications. *Annual Review of Statistics and Its Application* **10**.

SHU, Y., LIU, J., ZENG, X., HONG, H. G., LI, Y., ZHONG, H., MA, L. and FU, P. (2018). The effect of overhydration on mortality and technique failure among peritoneal dialysis patients: a systematic review and meta-analysis. *Blood purification* **46** 350–358.

STEVENS, E., ANTIGA, L. and VIEHMANN, T. (2020). *Deep Learning with PyTorch: Build, train, and tune neural networks using Python tools*. Manning.

TJANDRA, D., HE, Y. and WIENS, J. (2021). A Hierarchical Approach to Multi-Event Survival Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* **35** 591–599.

TSENG, J.-S., CHIANG, C.-J., CHEN, K.-C., ZHENG, Z.-R., YANG, T.-Y., LEE, W.-C., HSU, K.-H., HUANG, Y.-H., LIU, T.-W., HSIA, J.-Y. et al. (2022). Association of smoking with patient characteristics and outcomes in small cell lung carcinoma, 2011-2018. *JAMA network open* **5** e224830–e224830.

VAN STEENKISTE, S., CHANG, M., GREFF, K. and SCHMIDHUBER, J. (2018). Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *arXiv preprint arXiv:1802.10353*.

WANG, Q., FEDERICI, M. and VAN HOOF, H. (2025). Bridge the inference gaps of neural processes via expectation maximization. *arXiv preprint arXiv:2501.03264*.

XIE, J., GIRSHICK, R. and FARHADI, A. (2016). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning* 478–487. PMLR.

XU, J., KALBFLEISCH, J. D. and TAI, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* **66** 716–725.

YEGNANARAYANA, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

YUAN, Q., CAI, T., HONG, C., DU, M., JOHNSON, B. E., LANUTI, M., CAI, T. and CHRISTIANI, D. C. (2021). Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. *JAMA Network Open* **4** e2114723–e2114723.